DOCUMENT RESUME

ED 151 875 CS 502 011

AUTHOR Mancillas, William R. Todd; Kibler, Robert J.
TITLE A Test of Concurrent Validity for Knapp, Hart and Dennis' Linguistic Indices of Deception.
PUB DATE Nov 77
NOTE 36p.; Paper presented at the Annual Meeting of the Western Speech Communication Association (Phoenix, Arizona, November 18-21, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS College Students; Content Analysis; Language Research; Language Usage; Linguistics; *Measurement Techniques; *Speech Communication; *Validity
IDENTIFIERS *Deception

ABSTRACT

Previous research has demonstrated that 11 linguistic indices discriminate between truthful and deceptive language behavior. On the basis of subjective judgment, the original researchers arranged these indices in clusters (dimensions) of deceptive language behavior: uncertainty, vagueness, negative affect, and reticence. The present study sought to obtain concurrent validity for these indices by correlating them with other indices which have already been empirically validated as measures of various aspects of uncertainty, vagueness, negative affect, and reticence. Analysis of the data obtained significant concurrent validity only for those indices purportedly measuring reticence. Little or no support was obtained for the indices presumably measuring uncertainty, vagueness, and negative affect. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151875

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A TEST OF CONCURRENT VALIDITY FOR KNAPP, HART AND DENNIS' LINGUISTIC INDICES OF DECEPTION

By

William R. Todd Mancillas
Rutgers University

And

Robert J. Kibler
Florida State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William R. Todd Mancillas

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Doctor William R. Todd De Mancillas is a recent graduate of the Department of Speech-Communication at The Florida State University and is presently employed as an assistant professor in the Department of Human Communication at Rutgers University.

Doctor Robert J. Kibler is a professor of communication in the Department of Instructional Communication at The Florida State University.

Appreciation is extended to Ms. Judy Shank (Dade Community College, Miami, Florida) for her assistance in tabulating data and performing many other tedious clerical tasks necessary to the completion of this study.

Appreciation is also extended to Ms. Jo Ann Zsilavetz and Ms. Dolores Palazzo (Rutgers Department of Human Communication) for their help in proofing, typing, and reproducing copies of this manuscript.

CS502011

In recent years, considerable research has been done attempting to identify and refine physiological and nonverbal indices of deception[1]. By comparison, relatively few studies have attempted to identify and refine linguistic indices of deception[2]. In an attempt to increase our knowledge of linguistic indices of deception, Knapp, Hart, and Dennis have conducted an exploratory study, which identified 14 linguistic indices which significantly discriminated between truth telling and deceptive language[3].

The Knapp, et al. indices were arranged into five clusters as follows:

1. Two indices comprised an Uncertainty cluster. The confidence ratio index was computed by dividing the "total number of words spoken. . .by the total time the S and confederate (interviewer) spent in verbal interchange"[4]. The different words index was a count of the total number of distinctly different words spoken by the interviewee.

2. Four indices comprised the Vagueness cluster: factual statements, self-experience references, past references, and leveling terms. The factual statements index was a count of the number of interviewee "statements which remark about an existing person(s), object(s), or set(s) of conditions that are empirically verifiable by the senses"[5]. The self-experience references index was a count of the number of "remarks made by the S which

expound upon actual experiences he has had (or is having) or upon activities he has engaged in (or is engaging in)"[6]. The past references index was a count of past tense verbs uttered by the interviewee (e.g., was, were, has, had, etc.). The leveling terms index was a count of "linguistic markers" of "egregiously overgeneralized statements"[7]. Knapp, et al. explain that "such common leveling terms as 'every,' 'all,' 'none,' 'nobody,' etc. were designated as being central to our linguistic probes, probes which hopefully revealed the relative 'allness' of the statements made"[8].

3. Three indices comprised a Reticence cluster. The total words index was a count of the total number of words spoken by the interviewee. The message duration index was a measure of the total duration of interviewee speaking time, measured to the nearest second. The probes index was a count of the number of interviewer probes.

4. Three indices comprised a Dependence cluster. The self-references index was a count of the number of first person pronouns uttered by the interviewee. The self-interest index was a count of "statements of either a factual or evaluative character, which state the costs or benefist which have (or have not) accrued, are (or are not accruing, should (or should not) accrue to self"[9] The other references index was a count of the number of third person plural pronouns uttered by the interviewee.

5. Lastly, a Negative Affect cluster was comprised of two indices. The number of group references index was a count of the number of first person plural pronouns uttered by the interviewee. The disparaging statements index was a count of the number of interviewee statements "depicting unfavorable actions or statements of another person, group or institution"[10].

The specific findings of the Knapp, et al. study were that: (1) nondeceivers were more certain in their communication than deceivers; i.e., their confidence ratios were higher, and they used more different words than deceivers, (2) nondeceivers were clearer in their communication than deceivers; i.e., they uttered more factual statements, more self-experience statements, more past-references, and fewer leveling terms than deceivers, (3) nondeceivers were less reticent than deceivers; i.e., they uttered more total words and spent more time talking than deceivers, (4) nondeceivers were more independent than deceivers; i.e., they uttered more self-references, more self-interest statements, but fewer other references than deceivers, (5) nondeceivers were lower in negative affect than deceivers; i.e., they uttered more group references but fewer disparaging statements than deceivers.

For two reasons, the Knapp, et al. study is an important contribution to the deception literature. First, this study identified many more linguistic indices of deception than had been previously identified. Second, several

of the Knapp, et al. indices are highly similar to previously verified linguistic indices of deception. Specifically, the message duration index is similar to Matarazzo's duration of utterance index (operationally defined as the mean duration of interviewee utterances, excluding interviewer probes and pauses lasting longer than three seconds)[11]; the disparaging statements index is similar to Horvath's and Reid and Arther's complaint index (operationally defined as any interviewee complaints, including complaints about discomforting polygraph equipment, personal problems, and statements criticizing the polygraph recorder's ability to identify deceivers)[12]; and the total words and confidence ratio indices are similar to Mehrabian's speech output and speaking rate indices (operationally defined as the total number of word output and ratio of output to total amount of available speaking time)[13]. Thus, the Knapp, et al. findings demonstrate across study reliability for several indices.

Unfortunately, the Knapp, et al. study is flawed by a serious methodological problem. No concurrent validity checks were made to determine if, in fact, the indices measured what they were thought to measure. For instance, no empirical evidence was advanced in support of the assumption that "group indices" measure negative affect, nor was a logical argument advanced in support of this assumption. In the absence of concurrent validity, it is not possible to objectively determine the degree of accuracy in Knapp,

Hart, and Dennis' explanation that deceivers differ from truth tellers in that deceivers are more uncertain, vague, reticent, dependent, and higher in negative affect.

The present study was an attempt to assess the degree of concurrent validity for the Knapp, et al. indices. According to Cronbach and Meehl, concurrent validity is obtained for a given index when that index is demonstrated to correlate significantly with at least one other measure that is already known to measure aspects of the phenomenon of interest.[14] This procedure was implemented in the present study by computing correlations between the Knapp, et al. indices and five new ones which previous research has demonstrated to measure certain aspects of uncertainty, vagueness, negative affect, and reticence[15]. Specific a priori hypotheses follow.

Hypotheses 1 and 2—uncertainty indices: the silence quotient (SQ) will correlate positively and significantly with the total different words and confidence indices. Previous research has demonstrated that the SQ is a reliable and valid measure of interviewee output[16]. Positive correlations were expected between the SQ and uncertainty indices on the assumption that people talk more when they are certain of their position than when they are uncertain of their position.

Hypotheses 3 through 6--vagueness indices: The modified Flesch human interest score (MFHIS) will correlate positively and significantly with factual statements,

self-experience statements, and past reference statements, but negatively and significantly with leveling terms. Previous research has demonstrated that the MFHIS is a reliable and valid measure of readability, and that readability is inversely related to vagueness of a message[17]. Accordingly, the MFHIS was expected to correlate positively with factual, self-experience, and past reference statements, because Knapp, et al. posit these as inversely related to vagueness. Similarly, the MFHIS was expected to correlate negatively with leveling terms because Knapp, et al. posit leveling terms as positively related to vagueness.

Hypotheses 7 through 10—vagueness indices: the modified Flesch reading ease score (MFRES) will correlate positively and significantly with factual statements, self-experience statements, and past reference statements, but will correlate negatively and significantly with leveling terms. Like the MFHIS, the MFRES has also been demonstrated to reliably and validly measure readability[18]. Accordingly, the same rationale applies to support the hypothesized correlations between the MFRES and the vagueness indices as was previously advanced in support of the hypothesized correlations between the MFHIS and vagueness indices.

Hypotheses 11 and 12—negative affect indices: the discomfort relief quotient (DRQ) will correlate positively and significantly with disparaging statements, but negatively and significantly with group references. Previous research has demonstrated that the DRQ is a reliable and valid

measure of negative affect[19]. Accordingly, it was expected to correlate positively with the disparaging statements index which Knapp, et al. posited as positively related to negative affect. However, the DRQ was expected to correlate negatively with the group references index, because Knapp, et al. posit this index as negatively related to negative affect.

Hypotheses 13 through 15—reticence indices: the SQ will correlate positively and significantly with the total words and message duration indices, but negatively and significantly with the probes index. Knapp, et al. posited the total words and message duration indices as negatively related to reticence and the probes index as positively related to reticence.

Method

Subject Selection

Thirty-seven female volunteer subjects participated in all stages of this study. They were invited to participate after a pool of approximately 200 students were screened for their attitudes toward abortion. These students were enrolled in basic speech courses taught at Florida State University during Fall Quarter, 1975. All Ss agreed with the statement, "A woman should have the right to get an abortion if she wants it," and that they could speak in support of this position for at least two minutes.

Independent Variables

Three independent variables were used in this study. One was an instructional variable consisting of two levels,

an interview in which the S was instructed to persuade a female interviewer that she believed it was wrong for a woman to obtain an abortion (deception interview), and an interview in which the S attempted to persuade a different female interviewer that it was proper for a woman to obtain an abortion if she wanted one (truth telling interview).

The second independent variable consisted of the order in which the Ss underwent the interviews. Half of the Ss underwent the truth telling interview first and the deception interview second (order "A"). The remaining Ss underwent the deception interview first and the truth telling interview second (order "B").

The third independent variable consisted of two different interviewers.

Dependent Variables

Eleven of the dependent variables were selected from the Knapp, et al. study: total different words, confidence ratios, factual statements, self-experience statements, past reference statements, leveling terms, disparaging statements, group references, total words, message durations, and probes. Four additional dependent variables were selected because they could be used to test for the concurrent validity of the above indices. These additional variables were operationalized as follows. The silence quotient was calculated by dividing tht total amount of interviewee speaking time by the total amount of time available for her to speak.

The modified Flesch reading ease score was calculated by determining the average number of syllables uttered per 100 words. The modified Flesch human interest score was calculated by determining the average number of "personal words" spoken per 100 words, where personal words were operationally defined as names of people or animals, personal pronouns, or words with masculine or feminine natural gender. Lastly, the discomfort relief quotient was computed by ing the total number of "discomfort words" spoken by an interviewee by the total number of "relief words" spoken by the interviewee. Discomfort words were operationally defined as words manifesting negative affect (e.g., murder, hostility, terrible), and relief words were defined as words manifesting positive affect (e.g., happiness, kind, helpful).

Procedure

Fifty-eight female students were invited to participate in the study. Four students declined to participate. An additional 9 students failed to show up at preassigned times. And 8 students were eliminated from the study because they did not follow instructions.

Ss were assigned in advance by a table of random number to one of four possible interview order-by-interviewer sequences: truth telling interview first with Ms. C. H. as the interviewer; truth telling interview first with Ms. J. S. as the interviewer; deception interview first with Ms. C. H. as the interviewer; and deception interview first with Ms. J. S. as the interviewer.

Upon arriving at the experimental setting, each S was advised that she would particpate in two interviews. In one interview she would try to convince the female interviewer that a woman should have the right to have an abortion if she wants one. In another interview she would try to convince a different female interviewer that a woman should not have the right to get an abortion if she wants one. The S was also advised that a prize of $25.00 would be awarded to the participant who appeared most convincing in both interviews. S was then told whether the first interview would be deception or truth telling, and given four to five minutes to prepare for the interview.

At the end of the preparation period, S was led to another room for the interview. The interview was tape recorded, with the prior consent of S.

At the end of the interview, S completed a brief questionnaire indicating whether she had followd the truth telling/deception instructions. Eight Ss were eliminated from the study because their responses indicated that they had not told the truth during the truth telling interviews or had not lied during the deception interviews.

After completing the post-interview questionnaire, S was allowed four to five minutes to prepare for the second interview. S was then led to another room where the second interview took place.

After the second interview, S completed another questionnaire indicating whether she had followed the truth

telling/deception instructions, whether the money prize served as an incentive for lying and telling the truth during the appropriate interviews, and whether S had prior knowledge of the study. S was then thanked for her participation in the study and asked not to disclose information about it for at least two weeks thereafter.

The interviewers received specialized training prior to participating in the study. Interviewers were instructed to begin the interviews with the following question, "Do you think a woman should have the right to get an abortion if she wants one?" and not to ask additional questions unless the interviewee ran out of things to say. The interviewers were further instructed to allow at least a three second pause before asking a short, open-ended question (e.g., Do you have anything else to say? What else can you tell me about your feelings on this subject? Would you care to elaborate a bit further?). In short, the interviewers were instructed to be passive, neither encouraging nor discouraging participation, and only prompting S if it appeared that she was not going to talk anymore.

A pilot test helped internalize appropriate interviewer psychomotor behaviors. After six pilot Ss had been interviewed, the interviewers and principal investigator listened to tape recordings of the interviews. The investigator noted interactions between the interviewers and interviewees consistent with the above instructions and verbally reinforced them. Inappropriate interactions

were also noted (e.g., an interviewer probe elicited prior to the elapse of a three second interval), and appropriate emphasis was placed on the need to eliminate this behavior. A solution was reached by having the interviewers time pauses lasting three seconds with the aid of a stop watch. Then, without the aid of a stop watch, the interviewers were challenged to indicate when three seconds had elapsed. After this brief training period, the interviewers could signal a three second interval at will, with high accuracy and frequency.

Data Reduction and Statistical Analyses

The first 2 minutes of each interview were transcribed by three different people. More than one transcriber was needed to ensure reliable, valid transcriptions.[20] These transcriptions were compared to identify and resolve discrepancies. Very few discrepancies (fewer than .05%) were not resolved in this way.[21]

Transcriptions were punched on computer cards. The information punched on cards was identical to that on the original transcriptions, except that the interviewer probes were eliminated.

The cards were verified twice and then run through the Short Payne computer program. Output for this program consists of a reproduction of the original transcription, an alphabetized list of all utterances occurring in the transcription along with a frequency count for each utterance, a total count of the number of utterances occurring in the transcription, and a total count of the number

of different utterances occurring in each transcription.

A team of six graduate students worked as content analysis coders. The students were divided into teams of two members each, with the principal investigator serving as a third coder on each of the three teams.

Pursuant to a recommendation made by Budd, Thorp, and Donahew, coders were provided with written as well as oral instructions on how to compute the indices[22]. Holsti reliability coefficients were computed for 3 of the indices: factual statements (.95), self-experience statements (.98), and past-experience statements (.98).[23] The Holsti coefficient is a ratio of all possible coder agreements to coder disagreements. Pearson product moment correlation coefficients were computed for 6 of the indices: confidence ratios (1.00), disparaging statements (.96), number of leveling terms (.98), message durations (1.00), silence quotients (1.00), and discomfort relief quotients (.88).

It was not necessary to compute reliability coefficients for 6 of the indices (total number of words, total number of different words, total number of probes, total number of group references, modified Flesch reading ease scores, and modified Flesch Human interest scores), because each of these was computed using a manual or machine count of clearly defined phenomena (e.g., recognition of an utterance, recognition of an interviewer probe).[24] However, each count was taken at least twice by two different coders to avoid counting errors.

A brief note is needed explaining why two different reliability coefficients were computed. The Holsti coefficient was used when there was insufficient intercoder variability to warrant effective use of the Pearson product moment correlation coefficient. For instance, for nearly all subjects, coders computed a value of 0.00 on the self-experience statements index. Thus, there was near unanimous agreement among all coders that none of the subjects made references to past events. The intercoder agreement was high, and the intercoder reliability coefficients should reflect this. But because of the absence of appreciable intersubject variability, the resulting index would have been spuriously low had it been computed according to the Pearson product moment correlation[25]. Accordingly, the Holsti coefficient was computed in those instances where it would result in a more accurate indication of intercoder reliability.

Statistical Analyses

Using the Bio-Medical 08D computer program[26], a series of Pearson product moment correlations were computed in accordance with the previously specified a priori hypotheses. Also, each of these correlations was tested for significance using the t-test procedure recommended by Hays,[27] and each of these t-tests was conducted twice—once over the truth telling interview data and once over the deception interview data.

Each related cluster of hypotheses was tested at alpha levels derived from the Bonferroni method.[28] Specifically, an overall alpha level of .05 was assigned to each "cluster" of hypotheses, where a cluster was defined as all the hypotheses relating to the same dimension of deceptive language behavior proposed by Knapp, et al., namely the uncertainty, vagueness, negative affect, and reticence dimensions. The overall alpha level was divided by the number of specific hypotheses comprising a specific cluster to determine the particular alpha level at which each individual hypothesis was tested. This procedure resulted in specific alpha levels as follows: the two uncertainty hypotheses were each tested at alpha equal to .025 (.05/2 = .025); the eight vagueness hypotheses were each tested at alpha equal to .006 (.05/8 = .006); the two negative affect hypotheses were each tested at alpha equal to .025; the three reticence hypotheses were each tested at alpha equal to .017 (.05/3 = .017).

Results

Hypotheses 1 and 2—Uncertainty Indices

Truth telling interview results: silence quotients correlated significantly with total number of different words ($r = .75$) and confidence ratios ($r = 2.51$). The obtained t values were 6.73 and 2.51 respectively (at df = 35, alpha = .025, critical $t \cong 2.025$).

Decention interview results: silence quotients correlated significantly with total number of different

words ($r = .77$; $t = 7.12$). However, silence quotients did not correlate significantly with confidence ratios ($r = .10$; $t = .59$).

These results fully support hypothesis 1, but only partially support hypothesis 2.

Hypotheses 3 through 6—Vagueness Indices

Truth telling interview results: the modified Flesch human interest scores correlated significantly with factual statements ($r = .48$), but nonsignificantly with self-experience statements ($r = .25$) and past references ($r = .09$). The obtained t values were 3.23, 1.52, and .53 respectively (at $df = 35$, alpha = .006, critical $t \cong 2.75$). Since a positive correlation was obtained between the modified Flesch human interest scores and leveling terms ($r = .38$), a significance test was unnecessary, as hypothesis 6 had specified a negative—as opposed to positive—correlation between these indices.

Deception interview results: the modified Flesch human interest scores correlated nonsignificantly with factual statements ($r = .27$), past references ($r = .09$), and leveling terms ($r = -.28$). The obtained t values were 1.67, .53, and -1.73 respectively. Since a negative correlation was obtained between the modified Flesch human interest scores and self-experience statements ($r = -.09$), a significance test was unnecessary, as hypothesis 4 had specified a positive correlation between these indices.

The above results provide partial support for hypothesis 3, but no support for hypotheses 4 through 6.

Hypotheses 7 through 10—Vagueness Indices

Truth telling interview results: the modified Flesch reading ease scores correlated -.12 with factual statements, -.06 with self-experience statements, .02 with past references, and .04 with leveling terms. It was not necessary to conduct significance tests on the first two correlations, because hypotheses 7 and 8 specified positive—not negative—correlations. Similarly, it was not necessary to conduct a significance test on the positive correlation between modified Flesch reading ease scores and leveling terms, because hypothesis 10 specified a negative correlation. The .02 correlation with past references was nonsignificant ($t = .12 <$ critical $t = 2.75$: $df = 35$, alpha = .006).

Deception interview results: the modified Flesch reading case scores correlated -.19 with factual statements, -.27 with self-experience statements, and -.21 with past references. None of these correlations needed to be tested because hypotheses 7 through 9 specified positive correlations. Reading case scores correlated nonsignificantly with leveling terms ($r = -.12$, $t = -.72$).

These results fail to support any of the hypotheses 7 through 10.

Hypotheses 11 and 12—Negative Affect Indices

Truth telling interview results: discomfort relief quotients correlated nonsignificantly with total number of disparaging statements (r = .02) and total number of group references (r = -.21). The obtained t values were .12 and -1.30 respectively (at df = 35, alpha = .025, critical $t \cong 2.025$).

Deception interview results: discomfort relief quotients correlated nonsignificantly with total number of disparaging statements (r = .13; t = .78). The correlation between discomfort relief quotients and leveling terms was positive (r = .03) and contrary to the negative correlation predicted by hypothesis 12; accordingly, no significance test was conducted.

These findings fail to provide support for either hypothesis 11 or 12.

Hypotheses 13 through 15—Reticence Indices

Truth telling interview results: silence quotients correlated significantly with total number of words (r = .78), message durations (r = 1.00), and probes (r = -.71). The obtained t values were 7.44, 5.92, and -6.00 respectively (at df = 35, alpha = .017, critical $t \cong 2.457$).

Deception interview results: silence quotients correlated significantly with total number of words (r = .79), message durations (r = 1.00), and probes (r = -.49). The obtained t values were 7.67, 5.92, and -3.33 respectively.

The above findings fully support hypotheses 13 through 15.

Interpretation of Results

The results of this study demonstrate considerable concurrent validity for the three Knapp, et al. indices purportedly measuring reticence. Thus, it would appear that there is empirical support for Knapp, Hart, and Dennis' assumption that number of words spoken, message durations, and probes are valid indices of communication reticence.

However, no support was obtained for those hypotheses testing concurrent validity for the Knapp, et al. indices presumably measuring negative affect (disparaging statements and leveling terms),[29] nor was concurrent validity obtained for the Knapp, et al. indices purportedly measuring vagueness (factual statements, self-experience statements, past references, and leveling terms). Similarly, only slight concurrent validity was obtained for the Knapp, et al. indices purportedly measuring uncertainty (different words and confidence ratios). Accordingly, one is tentatively advised not to accept the Knapp, et al. assumptions that their indices, in fact, measure negative affect, vagueness, or uncertainty.

Discussion

There are three possible reasons why this study failed to obtain substantial concurrent validity support for the Knapp, et al. indices.

First, it is possible that coding instructions were different in this study than in the original Knapp, et al. study. Because of space limitations, the original Knapp, et al. article allotted only brief space for coding instructions and operational definitions. (The essence of those instructions and definitions have been presented in this article.) Although the present investigators attempted to implement those coding instructions as accurately as possible and feasible, it is nonetheless possible that one or more of the original indices may have been operationalized differently in the present study than in the original study.[30] Of course, if there were differences in measurement procedures, then these differences may have resulted in a failure to obtain the expected significant correlations between the original indices and those introduced in this study. However, owing to the care with which coding was conducted in the present study, and the additional information obtained through personal correspondence between the present and original investigators, it is thought that the possibility of disparity in measurement procedures is a relatively unlikely artifact affecting the results of this study.[31]

A second reason for failing to obtain concurrent validity for the Knapp, et al. indices may be attributable to the rigorous alpha levels implemented in this study. Four of the 15 hypotheses were tested at .025 alpha levels, 3 at .017 alpha levels, and the remaining 8 hypotheses at

.006 alpha levels.[32] The reason for implementing the Bonferroni method, which resulted in these unusual, conservative alpha levels, is that this is an effective method for reducing the problem of inflated alpha that occurs when many statistical tests are performed on the same data.[33] In fact, the overall alpha for this study is .18, or the equivalent overall alpha for experiment for any study in which four independent statistical tests are computed at the .05 level of significance.[34] By comparison, had each of this study's 15 hypotheses been tested individually at the .05 level of significance (unfortunately, not an uncommon practice in our field), the overall alpha level would have been .54, a level three times greater than the .18 level obtained by implementing the Bonferroni method.

It also needs to be pointed out that even if each of this study's 15 hypotheses had been tested separately at .05 alpha levels, the results would have been almost identical to the present findings. In point of fact, only one additional hypothesis would have been confirmed; specifically, the predicted negative correlation between modified Flesch human interest scores and leveling terms (obtained $t = -1.72$ critical $t = 1.69$, $df = 35$). Thus, the implementation of conservative alpha levels was not a cause for failure to obtain concurrent validity for the Knapp et al. indices.

A third reason for failing to obtain concurrent validity for the Knapp, et al. indices may be attributable

to sex differences between the subjects used in the original and present investigation. Knapp, et al. used male volunteers; the present investigators used female volunteers. Female subjects were used in the present study as part of an effort to determine whether the Knapp, et al. indices discriminated women's truth telling and deceptive language behaviors.[35] (Because of space limitations, it was not possible to discuss the methods and findings of this effort. Details concerning rationale, statistical analyses, and findings are presented in a companion paper.[36]) Since Knapp, Hart, and Dennis make no case for presuming that their indices would be more valid measures of men's than women's speech, one may presume that they intended for these indices to be valid measures of uncertainty, vagueness, negative affect, and reticence in speech, regardless of source gender. Accordingly, it would appear inadvisable to fault the concurrent validity procedures implemented in the present study merely because of differences in sex of subjects between the original and present investigation.

In summary, several possible artifacts were discussed which might have accounted for the present study's failure to obtain concurrent validity for the Knapp, et al. indices of uncertainty, vagueness, and reticence. But careful consideration of these possible artifacts led to the conclusion that none of these artifacts are sufficient to explain this study's nonsignificant results. Perhaps insufficient concurrent validity was obtained for the simple

reason that the Knapp, et al. indices do not measure what they were thought to measure. This possibility appears more likely when one gives particular attention to some of the indices and the underlying constructs they were presumed to measure. For instance, it is indeed difficult to understand why the index "group references" (operationally defined as the number of first person plural pronouns) would be expected to be a valid index of negative affect. The present investigators can think of nothing inherently hostile, disparaging, or negative in affect about the words "we" and "us", yet Knapp, Hart, and Dennis asserted that these were valid indices of negative affect. The present investigation was undertaken in an effort to determine the credibility of these assumptions by implementing systematic empirical investigation. And since the results of the present investigation offer only limited concurrent validity support for the Knapp, et al. indices, it is suggested that these indices are not the valid indices of uncertainty, vagueness, and negative affect originally presumed—but undemonstrated—by Knapp, Hart, and Dennis.

Future investigations might attempt to obtain concurrent validity for the Knapp, Hart, and Dennis indices by using different concurrent validity procedures. Or, perhaps, a rejoinder to the present study would help in further refining the methodological and conceptual issues raised herein.

Footnotes

[1]Studies attempting to identify and refine physiological indices would include: V. Appel, L. McCarron, and B. Manning, "The Eyeblink Rate: Behavioral Index of Threat," Journal of Counseling Psychology, 18 (1968), 153-157; R. J. Cutrow, A. Parks, N. Lucas, and K. Thomas, "The Objective Use of Multiple Physiological Indices in the Detection of Deception," Psychophysiology, 9 (1972), 578-588; H. J. Dana and C. C. Barnett, "The Emotional Stress Meter," in Academy Lectures on Lie Detection, ed. V. A. Leonard (Springfield, Ill.: Thomas, 1957); P. O. Davidson, "Validity of the Guilty-Knowledge Technique: The Effects of Motivation," Journal of Applied Psychology, 52 (1968), 62-65; D. A. Day and B. D. Rourke, "The Role of Attention in 'Lie-Detection,'" Canadian Journal of Behavioral Science, 6 (1974), 270-276; R. A. Dienstbier and P. O. Munter, "Cheating as a Function of the Labeling of Natural Arousal," Journal of Personality and Social Psychology, 17 (1971), 208-213; L. A. Gustafson and M. T. Orne, "The Effects of heightened Motivation on the Detection of Deception," Journal of Applied Psychology, 47 (1963), 408-411; Gustafson, L. A. and M. T. Orne, "The Effects of Task and Method of Stimulus Presentation on the Detection of Deception," Journal of Applied Psychology, 18 (1964), 383-387; S. Kugelmass and I. Lieblich, "Effects of Realistic Stress and Procedural Interference in Experimental Lie Detection," Journal of Applied Psychology,

50 (1966), 211-216; D. T. Lykken, "The GSR in the Detection of Guilt," Journal of Applied Psychology, 43 (1959), 385-388; D. T. Lykken, "The Validity of the Guilty-Knowledge Technique The Effects of Faking," Journal of Applied Psychology, 44 (1960), 258-262; R. I. Thackray and M. T. Orne, "A Comparison of Physiological Indices in Detection of Deception," Psychophysiology, 4 (1968), 329-339; and R. I. Thackray and M. T. Orne, "Effects of the Type of Stimulus Employed and the Level of Subject Awareness on the Detection of Deception," Journal of Applied Psychology, 52 (1968), 234-239.

Studies attempting to identify and refine nonverbal indices would include: Vanieta Atmyanandana, "An Experimental Study on the Detection of Deception in Cross-Cultural Communication," Diss. Florida State 1976; Paul Ekman and Wallace V. Friesen, "Detecting Deception from the Body or Face," Journal of Personality and Social Psychology, 29 (1974), 288; Eckhard H. Hess, "Your Pupils Always 'Tell' Truth, Give Away the Lies," Miami Herald, 31 July 1975, p. 9-A; A. R. Luria, "The Method of Recording Movements in Crime Detection," Zeitschrift fur Angewandte Psychologie, 35 (1930), 139-183; C. Jaffe, E. Millman, and B. Bormann, "An Attempt to Condition an Eyeblink Response to Verbal Deception," Psychological Reports, 12 (1966), 421-422; N. R. Maier, "Sensitivity to Attempts at Deception in an Interview Situation," Personnel Psychology, 19 (1966), 55-65; N. Maier and J. Janzen, "The Reliability of Reasons

Used in Making Judgments of Honesty and Dishonesty," Perceptual and Motor Skills, 25 (1967), 141-151; N. Maier and J. Thurber, "Accuracy of Judgments of Deception when the Interview is Watched, Heard, and Read," Personnel Psychology, 21 (1968), 23-30; Albert Mehrabian, "Nonverbal Betrayal of Feeling," Journal of Experimental Research in Personality, 5 (1971), 64-78; Scott Wilson, "Channel Differences in the Detection of Deception," Diss. Florida State 1975.

[2]To the present authors' knowledge, the following studies are the only ones which have been done attempting to refine linguistic indices of deception: F. S. Horvath, "Verbal and Nonverbal Clues to Truth and Deception During Polygraph Examinations," Journal of Police Science and Administration, 1 (1973), 138; Mark Knapp, Roderick P. Hart, and Harry S. Dennis, "An Exploration of Deception as a Communication Construct," Journal of Human Communication Research, 1 (1974), 15-29; J. D. Matarazzo, A. N. Wiens, R. H. Jackson, and T. S. Manaugh, "Interviewee Speech Behavior Under Conditions of Endogenously-Present and Exogenously Induced Motivational States," Journal of Clinical Psychology, 26 (1970), 141-148; Mehrabian; M. T. Motley, "Acoustic Correlates of Lies," Western Speech Journal, 38 (1974), 81-87.

[3]Knapp, Hart, and Dennis.

[4]Knapp, Hart, and Dennis, p. 21.

[5]Knapp, Hart, and Dennis, p. 19.

[6]Knapp, Hart, and Dennis, p. 19.

[7]Knapp, Hart, and Dennis, p. 21.

[8]Ibid, p. 21.

[9]Knapp, Hart, and Dennis, p. 20.

[10]Ibid, p. 20.

[11]Matarazzo, Wiens, Jackson, and Manaugh.

[12]Horvath; Reid and Arthur.

[13]Mehrabian.

[14]L. J. Cronbach and P. E. Meehl, "Construct Validity in Psychological Tests," Psychological Bulletin, 52 (1955), 281-302.

[15]No concurrent validation checks were made for the dependence indices, because the present researchers were unable to locate linguistic indices which had already been demonstrated to be valid indices of psychological dependence.

[16]G. F. Mahl, "Disturbances and Silences in the Patient's Speech in Psychotherapy," Journal of Abnormal and Social Psychology, 53 (1956), 1-15.

[17]The modified Flesch human interest score is nearly identical to the more widely known Flesch human interest score (FHIS). The FHIS is comprised of two parts: a weighted sum of "personal words" and a weighted sum of "personal sentences." Personal words are operationally defined as personal pronouns (e.g., I, me, you, we, us, etc.), words with masculine or feminine natural gender (e.g., man, woman, boy, father, etc.), and proper names (e.g., John, Mary, etc.). Personal sentences are operationally defined

as spoken sentences (sentences intended as quotes), questions, requests, sentences addressed directly to the receiver, exclamations, and grammatically incomplete sentences whose full meaning must be inferred from the context. (An exhaustive list of personal words and examples of personal sentences, along with explicit instructions and coding guides, are available from the authors upon request.)

The difference between the MFHIS and FHIS is that the former is calculated on the basis of personal words only, while the latter is computed on the basis of both personal words and personal sentences. For the purposes of this study, the MFHIS was the more appropriate index for two reasons: (1) only a few subjects uttered personal sentences, thereby, assuring extremely high and positive correlations between the two indices ($r \simeq 1.00$); and (2) the computational formula weights the sum of personal words by a multiplicand equal to 3.635, but weights the sum of personal sentences by a multiplicand of only .314, which is only 8% as large as the multiplicand used in obtaining a weighted sum of personal words. Consequently, to have used the FHIS rather than the MFHIS index would have been misleading, as it would have suggested that the interest scores were partially constituted of personal sentences when, in fact, there were too few personal sentences to have had measurable impact on any of the subjects' human interest scores. For the above reasons, the MFHIS

procedure was implemented in the present study, and was operationally defined as the average number of personal words spoken per 100 words.

[18] The MFRES was calculated by determining the average number of syllables uttered per 100 words. See J. W. Gibson, C. R. Gruner, R. J. Kibler, and F. J. Kelly, "A Quantitative examination of differences and similarities in written and spoken messages," Speech Monographs, 33 (1966), 444-451.

The reliability and validity of Flesch human interest and reading ease scores have been established through numerous empirical tests, including: J. Chall and H. Dial, "Predicting Listener Understanding and Interest in Newscasts," Educational Research Bulletin, 27 (1948), 141-153; 168; "Foreign News Written Over Heads of Readers," Editor and Publisher, (December 28, 1946), p. 28; Rudolph Flesch, "How to Write Copy that Will be Read," Advertising and Selling, March 1947, p. 113; Rudolph Flesch, "A New Readability Yardstick," Journal of Applied Psychology, 32 (1948), 221-233; A. S. Gilinsky, "How Valid is the Flesch Readability Formula?" American Psychologist, 3 (1948), 261; D. R. Murphy, "Test Proves Short Sentences and Words Get Best Readership," Printers Ink, 218 (1946), 61-64; D. R. Murphy, "How Plain Talk Increases Readership 43% to 66%," Printers Ink, 220 (1947), 35-37; S. S. Stevens and Geraldine Stone, "Psychological Writing, Easy and Hard," American Psychologist, 2 (1947), 230-235.

[19]J. Dollard and O. H. Mowrer, "A Method of Measuring Tension in Written Documents," Journal of Abnormal and Social Psychology, 42 (1947), 3-32.

[20]L. A. Gottschalk, C. N. Winget, and G. C. Gleser, Manual of Instruction for Using the Gottschalk-Gleser Content Analysis Scales: Anxiety, Hostility, and Social Alienation—Personal Disorganization (Berkeley: University of California Press, 1969).

[21]The transcription procedures used in this study were different from those commonly used in most content analysis studies. In many studies, a single transcriber transcribes all the verbal material of interest, but without the transcriber having been trained to a minimum criterion of accuracy. The assumption here is that there is no need to test for transcription accuracy, because the task is straightforward and unlikely to result in an inaccurate written duplication of the oral message. This is an undesireable procedure because: a) without training the transcriber to a minimum criterion of accuracy it becomes difficult, if not impossible, to determine the lower bound of transcription accuracy; and b) regardless of how straightforward the task may appear to the experimenter, it may be less clear to the transcriber how the material should be transcribed. Some sort of empirical check is needed to clarify the experimenter's instructions and to objectively determine the transcriber's ability to follow those instructions.

Another frequently used transcription procedure is to

train the transcriber to a minimum criterion of accuracy. While this procedure is better than the previously discussed procedure (at least one obtains an estimate of transcription reliability), it may nonetheless be faulted on the basis of two implicit assumptions. First, one assumes that the transcriber will maintain the minimum criterion level of accuracy throughout the transcription process. This assumption is in error because a single transcriber may not be able to maintain acceptable criterion for as long a period of time as may be needed. Transcribing oral speech is tedious work, and it is difficult to do tedious work without making occasional errors. Moreover, the probability of making errors increases substantially when one is required to considerable amounts of tedious work. Inasmuch as many content analysis studies require several hours of transcription activity (the present study required well over 150 hours to transcribe subject interviews), one must be concerned about the likelihood of transcription error due to tedium and stimulus fatigue.

A second assumption is that all the materials to be transcribed are no more difficult than those initially used in training the transcriber to a minimum criterion of accuracy. This assumption is questionable, because it is unlikely that all oral interactions are equally as difficult to transcribe. For instance, some people speak at lower tones than others. If the transcriber were not able to accurately transcribe low tones, this would neces-

sarily increase transcription error. The present study sought to avoid this and the above discussed problems by having each language sample transcribed by three different transcribers. This approach led to the identification and resolution of many transcription errors which would have otherwise gone undetected had only one transcriber been used.

[22]Richard W. Budd, Robert K. Thorp, and Lewis Donohew, Content Analysis of Communications (New York: The Macmillan Company, 1967).

[23]Ole R. Holsti, Content Analysis for the Social Sciences and Humanities (Reading: Addison Wesley Publishing Company, 1969).

[24]These clearly defined units are sometimes referred to as the "manifest" content of a message. See, for instance, Holsti, p. 12.

[25]Fred Williams, Reasoning with Statistics (New York: Holt, Rinehart, and Winston, Inc., 1968).

[26]W. J. Dixon (ed.), BMD: Biomedical Computer Programs (Berkeley, Calif.: University of California Press, 1974).

[27]William Hays, Statistics (New York: Holt, Rinehart, and Winston, 1963), p. 529.

[28]P. A. Games, "Multiple Comparisons of Means," American Educational Research Journal, 1 (1971), 531-565; R. G. Miller, Simultaneous Statistical Inference (New York: McGraw-Hill, 1966).

[29]Interestingly, separate statistical analyses indicated that the disparaging statements index was the only one of the original 14 indices which significantly discriminated between the present study's truth telling and deceptive language samples ($F = 20.94$; critical $F = 5.57$; alpha = .025; df = 1,33). The mean for truth telling language samples was .36; the mean for deception language samples was 2.29. Details concerning this and related analyses are discussed in a companion paper (See footnote 36 for complete reference).

[30]Copies of the written instructions prepared for content analysis coders are available upon request. It was not possible to reproduce them here because of their inordinate length, approximately sixty, single spaced pages.

[31]The present authors express sincere appreciation to Messrs. Mark Knapp and Roderick Hart for their assistance in clarifying certain sections of their article.

[32]The authors were not able to locate a "t table" which depicted precisely the t values needed to reject Ho at alpha levels equal to .017 and .006, given df = 35. Consequently, the authors relied on critical t values associated with alpha levels that closely approximated .017 or .006. Specifically, a critical t of 2.457 is associated with alpha equal to .01 (merely 7/100 smaller than the .017 level used in this study), and was thought precise enough to use for hypothesis testing. Similarly, a critical t of 2.75 is associated with alpha equal to .005 (merely

1/1000 smaller than the .006 level used in this study), and was thought precise enough to use for hypothesis testing. These t values were obtained from Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences (New York: McGraw-Hill, 1956), p. 248.

[33]Games; Miller.

[34]Overall alpha level computed on the basis of a formula recommended by Hays: $1 - (1 - \text{alpha})^K$, where K is the number of statistical tests conducted at a specified alpha. See Hays, p. 489.

[35]An attempted replication using female subjects is in keeping with recent arguments that behavioral scientists should devote more attention to communication differences between men and women. See, for instance, J. E. Baird, Sex Differences in Group Communication: A Review of Relevant Research," The Quarterly Journal of Speech, 62 (1976), 179-192: C. Kramer, "Women's Speech: Separate but Unequal?" The Quarterly Journal of Speech, 60 (1974), 14-24; Robert J. Kibler, Larry Barker, and Donald Cegala, "Effect of Sex on Comprehension and Retention," Speech Monographs, 37 (1970), 287-292; S. E. Purnell, "Sex Roles in Communication: Teaching and Research," Journal of the Western Speech Communication Association, 40 (1976), 111-120.

[36]William R. Todd Mancillas and Robert J. Kibler, "Linguistic Indices of Deception as Manifested by Women: A Content Analytic Study," Paper presented to the Organizational/Interpersonal Interest Group during the 1977 Western Speech Communication Association Convention (Phoenix, Arizona; November 18--23, 197